

A Short Note on Surveillance Testing

FD, CG, VK & NQ on behalf of LMH Governing Body

13/09/2020

Appropriate decisions about whether to continue or curtail in-person teaching and whether to expand or restrict access to communal spaces such as dining halls demand reliable information on the prevalence of infection among our students. This document outlines some simple strategies for tracking the number of students infected with coronavirus over the course of MT, under the assumption that Oxford lacks sufficient laboratory capacity to permit regular universal testing of the kind planned by the University of Cambridge. We focus on estimating coronavirus prevalence among students in college or university-provided residential accommodation. While it would be preferable to consider the entire university population, given limited testing and logistical capacity, a reasonable approach is to focus on the subpopulation in which transmission is most likely to occur. The problem of estimating infection prevalence is fundamentally different from that of determining whether a particular individual has the coronavirus. In particular, the high test specificity and sensitivity along with rapid turnaround that are crucial for individual medical diagnosis are not absolutely essential for surveillance testing. Our calculations suggest that relatively modest resources may be sufficient to obtain valuable information regarding the spread and prevalence of coronavirus among our students. We view this issue as separate from and parallel to the “Early Alert Testing” service. The ideas we discuss are neither novel nor original, and we do not propose a detailed plan for surveillance testing. Our goals are primarily to explain randomized and pooled testing strategies to a non-specialist audience, and to calculate the approximate number of tests required to provide useful information. For simplicity we begin by abstracting away the problem of false negatives and false positives. The final section discusses how these affect our analysis.

The Bottom Line

Testing a sample of approximately 1400 students chosen at random each week from among those in residential accommodation could provide useful estimates of the prevalence and spread of coronavirus among the student population at Oxford, allowing us to guide potential policy changes. Using a simple pooling strategy, this would likely require fewer than 300 laboratory tests per week (or roughly 40 per day). Because capacity cannot be diverted from the Early Alert Testing service, the university would need to obtain roughly 2400 laboratory tests from another source over the course of MT. These need not have a rapid turnaround time:

a delay of several days would still allow us to assess the spread of infection. Pooled testing reduces the number of laboratory tests required but does not affect the number of samples that need to be collected. To reduce the burden of collecting samples from 200 students daily, swabs could be self-administered under the guidance of a healthcare professional.¹

Randomized Testing

As explained above, we focus on estimating coronavirus prevalence among students in residential accommodation. Suppose that this sub-population contains approximately 20,000 individuals of whom an unknown number C have the coronavirus.² With a perfect test, we could learn C by testing all 20,000 students. But what if we only have the capacity to carry out n tests, where n is much smaller than 20,000? In this case we cannot learn C exactly, but by employing *randomized testing* we can produce an estimate that is sufficiently accurate to guide our mitigation policies. The value of randomized testing is widely acknowledged. It underlies both the [Coronavirus Infection Survey pilot](#) conducted by the ONS, and the [REACT-1](#) survey conducted by the Department of Health and Social Care. Both of these surveys are extremely valuable, but as they aim to be nationally representative, neither employs a sufficiently large sample size to track local trends. As such we propose that randomized surveillance testing be carried out within the university.

The most basic form of randomized testing relies on a *simple random sample*. To use this procedure, we place all 20,000 student id numbers into a hat, mix them up thoroughly, and then draw out n at random. These n students are sent for a coronavirus test. We then use the number of students in our sample who have coronavirus to calculate an estimate E of the number students in the population who have coronavirus. For example, if one student in a sample of 2000 tests positive, then we would estimate that there are 10 cases in the population of students. Repeating this process weekly would allow us to estimate how the prevalence of coronavirus changes over time.

Because students are chosen at random, the estimated number of cases E may not equal the true number of cases C . Purely by chance, we might draw a sample of students in which no one is infected. In this case E would equal zero: an underestimate of C . We could also be unlucky in the opposite direction and, purely by chance, draw a sample that contains everyone at Oxford who is infected. In this case E would be an overestimate of C . Both of these possibilities, however, are remote: E is highly likely to be close to C . With an appropriate sample size, an estimate based on random testing is reliable enough to tell us with high confidence whether the number of cases is, say, greater than 50 or fewer than 20. It can also be used to determine whether prevalence is increasing and, if so, how quickly.

¹This is the procedure used by the ONS in their [Coronavirus Infection Survey pilot](#). A recent study suggests that this procedure is no less accurate than having healthcare workers conduct the swabs directly: <https://www.medrxiv.org/content/10.1101/2020.04.11.20062372v1>.

²We arrived at the figure of 20,000 by assuming that 100% of undergraduate students and roughly 66% of postgraduates are housed in college or university-provided residential accommodation. For more precise student numbers, see <https://www.ox.ac.uk/about/facts-and-figures/student-numbers>.

Relative to the Early Alert Testing system, which tests only symptomatic individuals, a crucial advantage of randomized testing is that it will also detect pre-symptomatic and asymptomatic cases. This is important because, as explained in the [SAGE report of 3rd September](#), “asymptomatic transmission is a key risk in university settings.” Indeed, a recent study cited in the SAGE report finds that only 18% of those aged 0–19 and 22% of those aged 20–39 who are infected with coronavirus show symptoms of the disease.³ Even if these figures are substantial overestimates, it is likely that a considerable fraction of infections among our students will go undetected by the Early Alert Testing system. While students themselves are at low risk of serious complications from coronavirus, members of staff are at substantially higher risk. By the time the Early Alert system registers an increase in cases among members of staff, it may already be too late for mitigation policies, such as curtailing in-person instruction, to be effective. Even if their turn-around times were somewhat slower than those for tests conducted through the Early Alert system, randomized tests could still provide more timely information about the spread of infection, given the likely extent of asymptomatic transmission among our students.

Table 1: Expected number of samples before the first positive result under random sampling without replacement from a population of size 20,000.

#Cases (C)	Prevalence (%)	Expected #Samples
20	0.10	952
30	0.15	645
40	0.20	488
50	0.25	392
60	0.30	328
70	0.35	282
80	0.40	247
90	0.45	220
100	0.50	198

The key question in designing a randomized testing protocol is how large a sample size to use. Fortunately, there are several simple ways to provide an approximate answer. If there are C total cases out of 20,000 students in the university then, on average, we will need to draw approximately $20,000/C$ samples to detect our first positive result.⁴ Table 1 provides exact results over a range of values for C . This calculation provides a rough-and-ready method for showing that a proposed sample size is likely *too small* to be useful. For example, $n = 250$ is too small a sample size to provide reliable estimates when C is close to 20 because we would, on average, require close to a thousand samples to obtain our first positive result if this were the true prevalence. Based on recent results from [REACT-1](#), we can say with high confidence that between 0.16% and 0.41% of individuals in the 18-24 age group had the coronavirus at

³<https://arxiv.org/abs/2006.08471>

⁴The exact value is $(N + 1)/(C + 1)$ for a population of size N containing C positives.

the start of September. This translates to roughly 30-80 out of 20,000. To provide reasonable estimates for C within this range, we should not consider sample sizes below 500.

A more refined way to address the question of how many samples are needed is by calculating *how close* the estimated number of cases E will likely be to the true number of cases C for different values of n .⁵ Suppose that there are 60 cases within the university. This would represent a slight uptick from the most recent [REACT-1](#) estimate of coronavirus prevalence among 18-24 year olds. With a sample size of $n = 1000$ the estimated number of cases E will fall in the range $[20, 100]$, inclusive, with approximately 90% probability.⁶ With a sample size of $n = 1500$ this range narrows to $[26, 94]$, as seen from Figure 1.

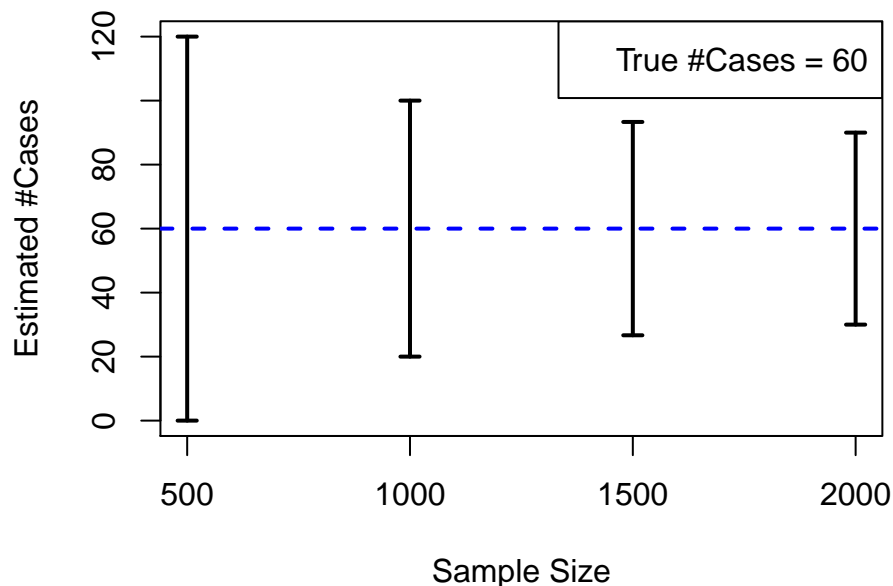


Figure 1: This figure shows how close the estimated number of cases will likely be to the true number of cases for different sample sizes. At a given sample size, the estimated number of cases has approximately a 90% chance of falling in the specified interval. See Footnotes 5 and 6 for further details.

A third way of deciding how many samples are needed is by calculating the chance that we will be able to reliably distinguish between two different levels of coronavirus prevalence given a particular sample size. Suppose, for example, that we initially believed there to be 50 cases in the university, corresponding to the current estimated prevalence among the university-aged population from [REACT-1](#). If cases increased to 100, we would have approximately a 68% chance of detecting an increase, with high confidence, based on a sample 1400 students.⁷

⁵The intervals presented in this paragraph and Figure 1 are calculated from the quantiles of a Hypergeometric(N, C, n) distribution where N is the population size, C is the number of positives in the population, and n is the sample size.

⁶Because E has a discrete support set, it is not possible to construct exact 90% intervals. The exact probabilities, to two decimal places, for the intervals with sample sizes (500, 1000, 1500, 2000) from Figure 1 and the text are (0.94, 0.88, 0.87, 0.87), respectively.

⁷These values correspond to a power calculation for an $\alpha = 0.1$ test of $H_0: p \leq p_0$ versus $H_1: p > p_0$ using the score test statistic $T_n = (\hat{p} - p_0)/SE(p_0)$ where $SE(p_0) = \sqrt{\kappa p_0(1 - p_0)/n}$ and $\kappa = (N - n)/(N - 1)$ is

If they increased to 200, we would have a 99% chance of detecting an increase. Roughly speaking, if cases were doubling weekly, then a sample of 200 students per day would be likely to detect an increase after one week, and nearly certain to do so after two weeks.

Broadly speaking, our calculations suggests that a simple random sample of approximately 200 students per day (1400 per week) would be useful to estimate population prevalence and determine whether, and how quickly, it may be increasing. This would amount to around half as many tests over the entire term as would be required for a *single* round of universal testing. The number of tests required could also be dramatically reduced by using a pooled testing strategy, as outlined below. While a simple random sampling procedure is the easiest to explain and to use for back-of-the-envelope calculations, there are several ways that it could be refined to increase the accuracy of our estimates. It is likely that coronavirus cases will tend to “cluster” within colleges, since the infection is most easily spread through frequent, close contact. If this is so, then a sampling procedure that allocates a certain number of tests to each college and draws a simple random sample *within* each college will be more efficient than a simple random sample taken at the university level. This is called a *stratified* random sample. Another possibility is to employ *adaptive sampling*, in which additional tests are allocated to colleges in which positive results are detected in a first batch of tests. While the analysis of such a design is more complicated mathematically, under assumptions that are likely to hold in the present example, it can make more efficient use of a limited number of tests.⁸ In a similar vein, one could consider a sampling procedure that exploits the fact that transmission is most likely to occur within household groups.

Pooled Testing

Pooled testing is a procedure that can be helpful in situations where laboratory capacity, rather than the ability to collect test samples, is the limiting factor in expanding testing. In this procedure, a single laboratory test is carried out on the *combined* samples of a group of individuals. For a sufficiently sensitive test, a negative result for the pooled sample indicates that no one in the group is infected. A positive test indicates that at least one individual in the group is infected, and it is then necessary to test the individual samples to pinpoint the infected individual or individuals. When the infection is sufficiently rare, pooled testing is highly efficient because virtually all pooled samples will test negative. For example, if $C = 50$ then pooling tests in groups of 5 would require approximately 4250 laboratory tests to achieve the equivalent of universal testing.⁹ This strategy has in fact been adopted by

a finite-population correction. In finite sample simulations, a standard normal critical value yields a slightly over-sized test, $\alpha \approx 0.13$ rather than 0.1, resulting in slightly higher power than this calculation would suggest.

⁸Thompson, S.K. (1990), “Adaptive Cluster Sampling”, *Journal of the American Statistical Association*, 85(412), pp. 1050-1059.

⁹These calculations, following Dorfman (1943), rely on two assumptions. First, we assume that tests are groups randomly to ensure statistical independence between samples in a given batch. Second, we assume that the test in question is sufficiently sensitive to ensure a positive test for the batch whenever at least one of the samples within it would have yielded a positive result when tested individually. Under these assumptions, if k is the batch size, n is the total number of samples, and p is the probability that an individual sample tests positive, the expected number of tests required to identify all positive samples equals $n[1/k + 1 - (1 - p)^k]$.

the University of Cambridge. In the absence of sufficient laboratory capacity, pooled testing could be combined with randomized testing. For example, our proposal to test a random sample of 1400 students each week would require approximately 300 laboratory tests each week. More sophisticated pooling strategies could potentially yield further improvement.¹⁰

What about false positives and negatives?

Thus far we have abstracted away the problem of false positives and false negatives. As our primary goal is to estimate prevalence rather than to diagnose particular individuals, this is not a major concern. First, if false positive and false negative rates are constant, we can still gain useful information by examining *changes* in estimated prevalence. Second, recent research has proposed estimates of these rates that can be used to correct estimates of population prevalence. More broadly, as we show in the appendix, false positives may be less of a problem among the university-aged population because this group is substantially more likely to be infected.

Appendix: Calculating False Positive Rates

The likelihood of a false positive depends on three factors: the *sensitivity* and *specificity* of a test along with the prevalence of coronavirus in the relevant population. Sensitivity is the share of people who test *positive* among those who *have coronavirus*. Estimates of the sensitivity of PCR tests vary, but range between 71% and 98%.¹¹ Specificity is the share of people who test *negative* among those who *do not have coronavirus*. We can bound the specificity of PCR tests fairly accurately using results from the [ONS Infection survey pilot](#). Between 1 June and 12 July, only 50 of the 122,776 samples collected as part of this survey tested positive. To compute a lower bound for the specificity, suppose that all 50 positive results were false positives. If true, this would mean that the ONS obtained 122,726 negative test results in a group of 122,776 people who did not in fact have coronavirus, implying a specificity of 99.96%. If any of the positive test results were true positives, then the true specificity must be higher. To be conservative, assume a sensitivity of 71% and a specificity of 99.96%. According to [REACT-1](#), the prevalence of coronavirus in the population as a whole was approximately 0.13% at the beginning of September. At this rate of prevalence, no more than 30% of positive test results will be false positives.¹² Among university-aged individuals, however, estimated prevalence was approximately 0.25%. Thus, among students no more than 18% of positives will be false positives. Whether this should be viewed as a high or low number, of course, depends on one’s perspective. As explained above, a policy of

¹⁰For a population prevalence of 0.25%, the optimal pool size under Dorfman’s assumptions is 21, which reduces the amount of laboratory tests required by 90% in expectation. To be conservative, we use a batch size of 5, yielding more modest savings of 79%. Much more sophisticated pooling strategies than Dorfman’s have been developed in the intervening years. For a recent refinement, see [Shental et al. \(2020\)](#).

¹¹<https://www.bmj.com/content/369/bmj.m1808>

¹²Defining $B = (100\% - \text{Specificity}) \times (100\% - \text{Prevalence}) / (\text{Sensitivity} \times \text{Prevalence})$, the false positive rate is given by $[B / (1 + B)]\%$.

testing only symptomatic individuals may miss up to 80% of all coronavirus cases among our students. Under a policy of universal testing, on the other hand, up to 18% of students instructed to self-isolate may have been inconvenienced unnecessarily.